Le poids des liens proches – Étude de la dynamique d'un grand réseau social

Stéphane Raux, Christophe Prieur

Le poids des liens proches : étude de la dynamique d'un grand réseau social

Stéphane Raux[*], Christophe Prieur[†]

Résumé

Les grands réseaux sociaux sont des réseaux animés dans lesquels les relations se nouent et se dénouent au fil du temps. Nous analysons ces dynamiques pour le réseau des commentaires du site Flickr à travers l'évolution de la structure globale du réseau, mais aussi celle des voisinages locaux et la manière dont ils se constituent.

Nous montrons que la structure générale du réseau est remarquablement stable sur la durée, et que l'essentiel de l'activité concerne des « liens proches ». Nous nous appuyons sur cette notion pour analyser les configurations locales et montrer les répercussions de ces liens sur la constitution du voisinage : les individus qui ont le plus de voisins choisissent ceux-ci parmi les voisins de leurs voisins, et ils se distinguent par leur activité très intensive, avec très peu de périodes d'interruption.

Introduction

En quelques années, l'analyse des réseaux sociaux est passée d'un courant de recherche en sciences sociales [Sco92, DF94] à un outil central dans le développement des plus récentes plateformes du web. Celles-ci s'appuient sur les interactions de leurs utilisateurs pour proposer des services qui couvrent aujourd'hui de nombreux domaines allant de l'organisation de l'information au commerce en ligne, en passant par la publications de photographies ou des services de rencontre ou de divertissement.

Le succès de ces plateformes, le volume très important de données qu'elles rassemblent et leur capacité à les organiser en réseau en font des sujets de recherche privilégiés [Boy04, CP08], tout comme le Word-Wide Web lui-même dans la décennie précédente, où les travaux autour de la problématique de la recherche d'information sur ce vaste corpus documentaire [ERC+00] ont abouti non seulement à la mise au point du *PageRank* de Google [BP98] mais également à la création d'un nouveau champ de recherche à partir de la découverte des propriétés que partagent un grand nombre de grand réseaux issus de disciplines

[*]Lip6, Université Pierre et Marie Curie et Liafa.
[†]Liafa, Université Paris-Diderot.

aussi variées que la biologie, la linguistique ou l'économie (et bien entendu l'informatique et la sociologie), grands réseaux qu'on appelle en anglais *complex networks* ou *small worlds* et en français grands réseaux d'interactions, petits mondes, ou graphes de terrains [BA99, WS98, New03, BS03].

La possibilité d'accéder à des données horodatées et le caractère très volatile de certains types de réseaux (comme les réseaux ad-hoc, pair à pair ou certains réseaux de communication) a aussi permi l'essor de l'étude des dynamiques des grands réseaux. Cette étude peut avoir des objectifs différents : on peut s'intéresser aux mécanismes dynamiques qui interviennent dans le réseau, ce qui débouche sur des champs d'applications comme l'épidémiologie ou le marketing viral [WD07]. On peut aussi s'intéresser à l'évolution de la structure des graphes et des groupes qui les composent [PBV07], ou encore essayer de modéliser la constitution de réseaux sociaux [JGN01].

La plupart de ces études s'intéressent surtout aux caractéristiques globales des réseaux et à l'identifications de groupes, mais on ne peut imaginer décrire avec finesse l'évolution d'ensemble d'un réseau sans porter le regard sur les innombrables dynamiques interindividuelles qui animent les acteurs de ce réseau. L'individu pris dans des chaînes d'interdépendances, pour citer pompeusement Norbert Elias [Eli91], n'est pas seulement l'atome constitutif du réseau, il en est aussi sa raison d'être. Comprendre comment l'individu peut agir malgré ou grâce aux liens qui l'entourent occupe depuis longtemps les sociologues. Tant la force des liens faibles de Granovetter [Gra78] que les trous structuraux de Burt [Bur92] montrent que le réseau peut être utilisé et ses liens mobilisés. Dans une perspective moins utilitariste et plus descriptive, toutes les recherches sur les réseaux égocentrés s'attachent à montrer la grande diversité des formes de sociabilité au travers des combinaisons de liens entre un individu, *ego*, et son entourage [Wel93, Gri98, PSS09].

Notre travail a pour objectif de mieux comprendre les mécanismes de constitution des liens et d'évolution de la structure d'un réseau, en adoptant deux échelles d'analyse : d'une part celle du graphe dans son ensemble, en étudiant l'évolution de ses principales caractéristiques, et d'autre part celle des sommets du graphe, en s'intéressant en particulier à la structure et aux modalités de formation de leur voisinage.

L'étude présentée ici porte sur l'évolution du réseau des commentaires du site Flickr.com, un site de partage de photos et de vidéos qui compte parmi les sites phares de ce « nouveau web » évoqué plus haut, dit « Web 2.0 » [O'R05]. Lancé en février 2004, Flickr a rencontré un très grand succès, qui lui vaut d'être racheté par Yahoo ! en mars 2005. Le réseau formé par ses utilisateurs a déjà fait l'objet de plusieurs études, qu'il s'agisse d'analyser les usages de ses fonctionnalités [PCB+09] ou d'étudier la structure générale et son évolution [KNT06].

Nous consacrons la première partie de cet article à la présentation des données et des méthodes que nous avons mises en œuvre pour leur analyse. Nous étudions ensuite dans la deuxième partie la structure générale du graphe, en comparant les résultats obtenus avec différentes méthodes de construction de

notre graphe. Nous constatons qu'après une courte période de mise en place, la struture est remarquablement stable, quelle que soit la méthode choisie. Nous mettons aussi en évidence l'importance des relations de courte distance au sein du graphe, ce qui nous amène à étudier dans la troisième partie la manière dont les sommets construisent leur entourage. Nous y montrons l'importance de l'origine des nouveaux voisins en distinguant le cas des voisins choisis parmi les « voisins de voisins » et en montrant les conséquences que ces choix peuvent avoir sur la structure du réseau local. Enfin, la quatrième partie est consacrée à la mesure de l'activité des sommets au fil du temps, et montre que la différenciation la plus pertinente ne concerne pas la durée totale d'activité, mais son intensité, en distinguant les sommets qui ont une activité continue de ceux qui ont une activité occasionnelle.

1 Données, modèles et outils

1.1 La base de commentaires Flickr

Nous avons travaillé sur un jeu de données obtenu en utilisant l'API publique de Flickr en août 2006. Ces données représentent plus de 500 millions de photographies et plus de 5 millions de membres.

Notre travail porte uniquement sur les commentaires de cette base de données. Un commentaire est toujours écrit par un utilisteur, que nous apellerons *émetteur*, sur la photographie d'un utilisateur que nous appellerons *destinataire*. Les messages ainsi enregistrés constituent un fil de discussion qui s'affiche en dessous de la photographie, ce qui permet aux utilisateurs d'interagir au sujet de la photographie. L'émetteur et le destinataire peuvent donc être une même personne, si l'auteur d'une photographie choisit de répondre aux commentaires déposés sur celle-ci.

Nous avons simpifié la structure de graphe biparti reliant des utilisateurs à des photographies pour nous concentrer sur les relations entre les individus : nous n'avons conservé que l'identifiant de l'émetteur, l'identifiant du destinataire et le *timestamp* correspondant au moment où le commentaire a été écrit, à la seconde près. Nous avons par ailleurs retiré les commentaires dans lesquels l'émetteur et le destinataires sont la même personne.

Nous obtenons ainsi une base de 39 594 157 commentaires qui ont été rédigés entre mars 2004 et juillet 2006 par 910 454 utilisateurs.

1.2 Formalisation

Les informations issues de la base de commentaires se présentent donc sous la forme d'une liste de liens dirigés classée par ordre chronologique de la forme (e, d, t), où e est l'identifiant de l'émetteur du commentaire, d celui du destinataire, et t le *timestamp* du commentaire.

Pour prendre en compte la dimension dynamique de nos données, on définit un intervalle de temps discret $T = [t_0, t_{max}]$, et pour tout $t \in T$ le graphe

non orienté $G_t = (V, E_t)$, où V est l'ensemble des nœuds et E_t l'ensemble des relations entre deux nœuds (u, v) qui « existent » à l'instant t. On peut adopter différentes stratégies pour déterminer si une relation existe ou non. Nous considérerons dans cet article que les commentaires sont cumulatifs : un lien entre u et v est considéré comme existant à l'instant t si les deux nœuds ont déjà échangé au moins un commentaire à un instant $t' \leq t$. Il est aussi possible de limiter cette condition en considérant que les relations disparaissent si aucun commentaire n'est réémis entre deux nœuds au-delà d'un intervalle de temps donné. Nous étudierons ce cas dans la partie 2.2.

Nous définissons par ailleurs le graphe $G = (V, E)$, qui contient l'ensemble des commentaires. Plus précisément :

$$E = \bigcup_{t \in T} E_t$$

Le *voisinage* $N_t(u)$ dans G_t (resp. $N(u)$ dans G) d'un nœud u est l'ensemble de nœuds v tels que $(u, v) \in E_t$ (resp. $(u, v) \in E$)). Le *degré* $\deg_t(n)$ dans G_t (resp. $\deg(u)$ dans G), est le nombre d'éléments dans $N_t(n)$ (resp. dans $N(n)$). La *distance* entre deux sommets u et v, notée $\text{dist}_t(u, v)$ est la longueur du plus court chemin (suite d'arêtes adjacentes) entre u et v dans G_t. S'il n'existe pas de chemin, on notera $\text{dist}_t(u, v) = \infty$. Le *voisinage à distance 2*, noté $N_t^2(u)$ dans G_t (resp. $N^2(u)$ dans G) d'un sommet u désignera l'ensemble des sommets qui sont des voisins de voisins de u sans être compris dans $N_t(u)$ (resp. $N(u)$).

Étant donnés deux sommets u et v voisins dans G, on appelle *écart entre u et v*, qu'on note $\widehat{u-v}$, la distance qui les sépare au moment où ils entrent en contact la première fois, *i.e.* $\widehat{u-v} = \text{dist}_t(u, v)$, où t est le plus petit entier de T tel que $\text{dist}_{t+1}(u, v) = 1$ et $\text{dist}_t(u, v) > 1$.

Un voisin v d'un sommet u est dit *proche* si l'écart entre eux vaut 2, et *lointain* sinon. Que v soit proche de u signifie que v était voisin d'un voisin de u avant qu'ils n'entrent en contact. Étant donné un sommet u, on notera $P(u)$ la proportion de voisins proches de u :

$$P(u) = \frac{\text{Nombre de voisins proches de } u}{\text{Nombre de voisins de } u}$$

1.3 Mesures

Nous utilisons quelques mesures courantes pour l'étude des grands réseaux :

Une *composante connexe* est un ensemble de sommets qui sont tous connectés entre eux par au moins un chemin. Leur identification est surtout intéressante pour identifier la composante connexe principale : les graphes de terrain se caractérisent en effet par l'existence d'une très grande composante connexe qui contient une grande proportion des sommets, puis d'un grand nombre de composantes connexes contenant très peu de sommets. En pratique, la composition des composantes connexes peut se calculer à la volée au moment du chargement du graphe au moyen d'un algorithme *union-find*.

On définit la *périphérie* d'un graphe comme l'ensemble des sous-graphes induits qui ne contiennent aucun cycle. Cette mesure est très utilisée en analyse de réseaux sociaux, en particulier dans le cas de la périphérie de la composante connexe principale, car elle identifie des sommets qui ont une position « marginale », mais qui restent malgré tout connectés à une grande partie du graphe (voir [Sei83]). La périphérie peut se calculer en temps linéaire au moyen d'un parcours en largeur partant des sommets de degré 1 (qui font nécessairement partie de la périphérie) et ne visitant que les noeuds qui n'ont qu'un seul sommet non encore visité.

Le nombre de *triangles* dans le graphe (*i.e.* de triplets de sommets tous trois connectés entre eux) permet de mesurer les densités locales dans G, en particulier le *coefficient de clustering*, qui désigne la probabilité pour que deux voisins u et v d'un même sommet soient également en relation. On parcourt pour chaque sommet l'ensemble de ses voisins et on compte leurs contacts communs. Ce calcul se fait en un temps très rapide en pratique (voir [Lat08] pour un survol des méthodes de calcul des triangles).

1.4 Méthodes

Ces mesures peuvent être effectuées directement sur le graphe G, ce qui revient à mesurer l'état de G_t à la fin de la période T si on considère le cas où les liens sont cumulatifs. Les mesures peuvent porter sur l'ensemble du graphe : taille et nombre de composantes connexes, taille de la périphérie ou coefficient de clustering. Elles peuvent aussi s'appliquer à chaque sommet u, par exemple la taille du voisinage ($N(u)$ comme $N^2(u)$). Les informations qui sont ainsi recueillies permettent ensuite d'effectuer des recoupements et de mieux comprendre les variations de profils entre les différents sommets.

Pour mesurer l'évolution de ces mesures au fil du temps, il faut effectuer les calculs au fur et à mesure de la lecture des commentaires. En terme d'implémentation, on charge alors l'ensemble des liens de G en mémoire sans indiquer de temps, puis on parcourt la liste des commentaires en mémorisant pour chaque lien la date de sa dernière occurrence. Dans le cas cumulatif, on teste la présence d'un lien dans G_t en vérifiant si sa dernière occurrence est inférieure ou égale à l'instant t.

On affiche ensuite l'état de G_t à intervalles réguliers : si les mesures se calculent en temps linéaire avec une faible constante et qu'on choisit des intervalles de quelques jours, on peut simplement utiliser les algorithmes des graphes statiques sur G_t lors de chaque intervalle. En revanche, on ne peut pas calculer l'évolution du coefficient de clustering avec cette méthode, car il serait trop coûteux d'effectuer la mesure à de nombreuses reprises. On utilise alors un algorithme dynamique qui fait le décompte des nouveaux triangles à chaque fois qu'une nouvelle relation apparaît dans le graphe.

Ces mesures peuvent être calculées pour mesurer l'évolution de l'ensemble du graphe, mais on peut aussi choisir de suivre un sommet u et d'étudier l'évolution de ses caractéristiques. Cette deuxième méthode permet d'utiliser des méthodes plus coûteuses en temps car on peut alors limiter les calculs aux commentaires

dans lesquels u apparaît.

Pour implémenter efficacement ces mesures dynamiques, il faut réduire au minimum les traitements de réinitialisation qui peuvent être nécessaires entre deux calculs. En fonction du type de mesure, on peut effectuer ces réinitialisations à la volée, lors de chaque étape de calcul ou bien on peut utiliser une pile pour ne réinitialiser que les valeurs qui ont été modifiées à l'étape précédente.

2 Dynamiques à l'échelle du graphe

2.1 L'évolution de l'activité

Une première approche consiste à évaluer l'évolution de l'activité des commentaires : pour chaque journée, nous avons mesuré le nombre de messages émis, le nombre d'utilisateurs différents qui ont émis un message et le nombre de destinataires différents auxquels les messages ont été adressés. Un utilisateur est considéré comme « actif » si il a émis ou reçu au moins un message au cours d'une journée. L'évolution du nombre d'utilisateurs actifs est représentée sur la figure 1. L'activité est faible jusqu'en janvier 2005, ce qui correspond à une

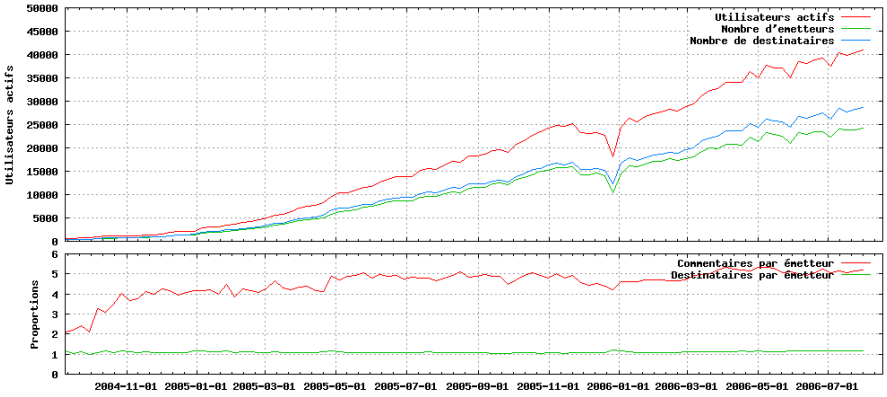


FIG. 1 – Évolution de l'activité sur l'ensemble de la période.

période de mise en place pendant laquelle le nombre moyen de messages par émetteur double, en passant de 2 à 4. L'activité croît ensuite de façon régulière tout au long de la période, à l'exception d'un décrochement au tournant de l'année 2005–2006. Le nombre de commentaires émis par rapport au nombre d'émetteurs augmente peu après la phase de mise en placen, il se stabilise rapidement autour de 5 commentaires par émetteur et reste constant ensuite.

Le rapport entre le nombre de destinataires et le nombre d'émetteurs est lui aussi constant : il y a en moyenne 1,1 destinataires par émetteur, ce qui suggère que les utilisateurs ont tendance à concentrer leurs commentaires sur quelques utilisateurs et à se répondre entre eux, dans un contexte de discussion.

La constance de ces indicateurs montre que l'augmentation de l'activité des commentaires est liée uniquement à l'augmentation du volume des utilisateurs. Si l'on ne tient pas compte de la période de mise en place, ceux-ci ne sont pas plus prolifiques à la fin de la période d'observation que lors du lancement du service : il y a un peu plus de 1 500 émetteurs par jour au début du mois de janvier 2005 et ils sont environ 20 fois plus nombreux à la fin du mois de juillet 2006, avec près de 30 000 émetteurs par jour.

Sans surprise, l'activité des utilisateurs varie en suivant des cycles d'une semaine. On observe un creux d'activité le dimanche, puis une reprise progressive de l'activité le lundi et un pic d'activité le mardi qui est suivi par une baisse progressive jusqu'à la fin de la semaine (cf. figure 2).

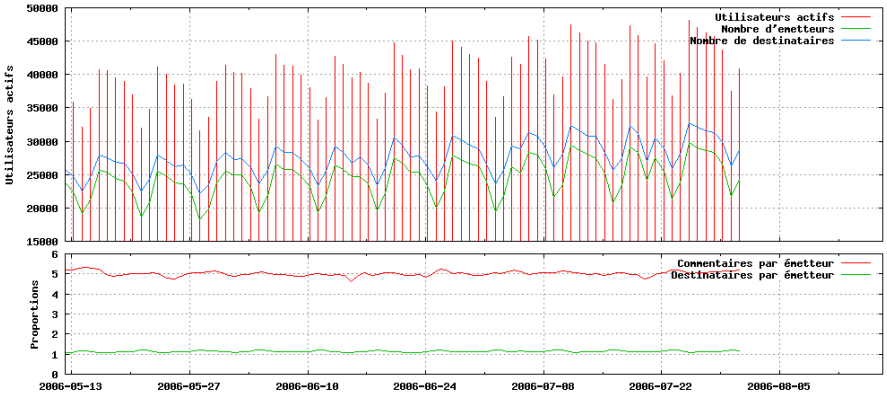


FIG. 2 – Évolution de l'activité pour les 3 derniers mois d'observation.

Ce premier survol permet d'observer deux phénomènes : l'élargissement du volume des commentaires, qui va de pair avec la popularité croissante du service, et la constance du comportement moyen des utilisateurs actifs. Pour comprendre comment cette activité se traduit en terme de relations, nous avons mesuré la structure générale du graphe et son évolution.

2.2 Structure générale du graphe

L'évolution de la taille de la composante connexe principale donne une bonne idée de l'accroissement de la taille du graphe car elle regroupe une grande partie de ses sommets. Nous avons d'abord construit notre graphe en adoptant le modèle cumulatif : une relation existe à partir du moment où au moins un commentaire a été émis. On peut cependant considérer que dans un contexte dynamique, les relations perdent de leur importance si elles ne sont pas réactivées régulièrement. Le simple fait d'émettre un commentaire est un engagement peu coûteux pour un utilisateur, la relation ainsi créée doit être renouvelée régulièrement pour être significative.

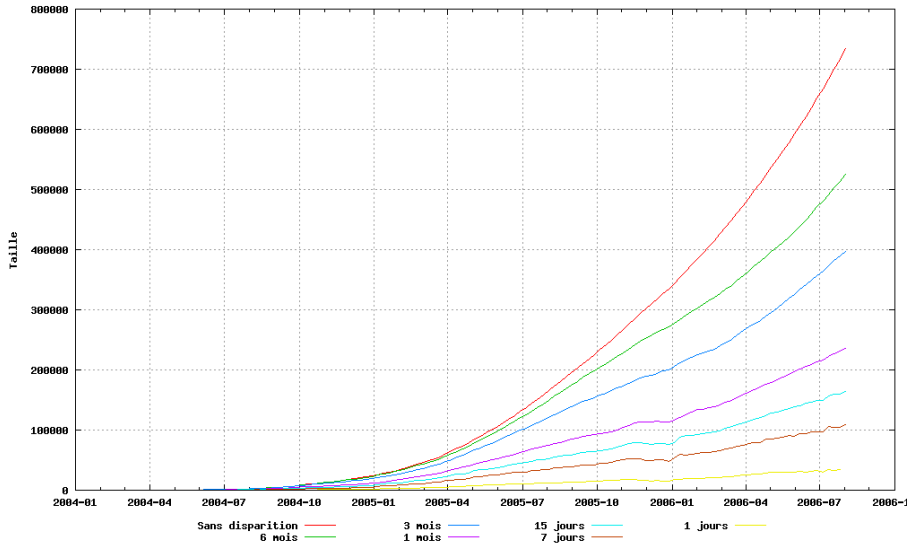


FIG. 3 – Évolution de la taille de la composante principale en fonction de la méthode de construction du graphe.

On peut alors modifier le modèle de construction de G_t en fixant un délai d'activité au delà duquel une relation est considérée comme abandonnée. Si à un moment donné la date de dernière activation d'une relation dépasse le délai fixé, on la retire du graphe. Par extension, on considère qu'un sommet u est actif dans G_t s'il est relié à au moins un autre sommet par une relation active au moment de la mesure, *i.e.* si $\deg_t(u) \geq 1$.

La figure 3 a été obtenue en utilisant la méthode de mesure dynamique de G_t telle que nous l'avons décrite dans la partie 1.4. La courbe la plus élevée correspond à l'accroissement de la composante connexe principale dans le cas où les liens s'ajoutent au fil des commentaires, les courbes inférieures ont été obtenues en choisissant différents délais d'activité, de six mois à une journée. La réduction du délai entraine une réduction de la taille de la composante connexe principale, mais celle-ci augmente quel que soit le délai choisi. Cela n'est pas surprenant puisqu'on sait déjà que l'activité quotidienne augmente tout au long de la période.

On peut ajouter à cela une contrainte de réciprocité dans les relations : on peut décider de ne conserver une relation entre deux sommets que s'il existe au moins un lien dans chaque direction pour cette relation, autrement dit si chacun des sommets a été au moins une fois émetteur et récepteur. On élimine ainsi les relations unilatérales qui peuvent être considérées comme de moindre importance si l'on s'intéresse aux interactions entre les utilisateurs. Le graphe des commentaires réciproques contient 259 395 sommets pour 25 818 794 commentaires. Cela signifie que les relations entretenues par moins d'un tiers $(28,4\%)$ des utilisateurs représentent près de deux tiers $(65,2\%)$ de l'ensemble des com-

mentaires. Cette dissymétrie nous montre qu'il existe au sein du graphe une minorité d'utilisateurs dont l'activité interne constitue une part importante de l'ensemble des commentaires.

Pour comparer les résultats obtenus avec les différentes méthodes de construction du graphe, nous rapportons les valeurs brutes obtenues pour la taille de la composante connexe principale au nombre de sommets actifs dans G_t. La figure 4 représente ainsi l'évolution de la taille relative de la composante connexe principale : les courbes de gauche concernent les mesures effectuées sur l'ensemble des commentaires, celles de droite correspondent aux commentaires réciproques. À titre de comparaison nous avons mesuré dans la figure 5 l'évolution de la taille relative de la périphérie de G_t.

On constate que les proportions sont très comparables, quels que soient le délai de suppression ou le graphe considéré. Dans tous les cas, on retrouve l'évolution décrite dans l'article de Kumar, Novak et Tomkins [KNT06] : une première période de mise en place de la structure du graphe, puis une stabilisation des proportions à partir de décembre 2004. La taille relative de la composante principale est alors en légère augmentation et reste proche de 85% des sommets actifs, la taille relative de la périphérie décroît légèrement en restant supérieure à 40%. Même en adoptant des délais de suppression très courts, de l'ordre de la semaine ou de la journée, ces proportions restent relativement stables, une fois passée la première période de constitution de la structure du graphe.

Ces résultats confortent nos premières observations sur l'activité. Ils montrent aussi que la structure relationnelle de notre graphe est remarquablement stable au fil du temps et qu'elle résiste bien aux contraintes de délai et de réciprocité des relations que l'on peut introduire dans la construction du graphe. Comme les commentaires réciproques représentent près des deux tiers de l'ensemble des commentaires, nous pouvons imaginer que l'intensité des relations entre ces sommets joue un rôle prépondérant dans l'élaboration de la structure du graphe : les commentaires non réciproques viennent grossir la taille du graphe mais ne modifient pas sensiblement sa structure.

2.3 L'importance des phénomènes locaux

Le rôle important des relations réciproques nous montre que les utilisateurs ont tendance à échanger plus de messages avec les personnes avec qui ils sont déjà en contact. Pour évaluer l'importance de ces phénomènes locaux dans la structure du graphe, on mesure pour l'ensemble des commentaires la proportion de liens répétés et de nouvelles relations. Lorsqu'une nouvelle relation est observée, on mesure son écart, en distinguant les nouvelles relations avec un voisin proche (écart de 2) des nouvelles relations avec un voisin lointain (écart supérieur ou égal 3). Nous considérons en effet que les membres d'un réseau ont une vision réduite de leur entourage. Ils peuvent avoir conscience d'une partie des personnes qui se situent dans leur entourage à distance 2, mais de leur point de vue il n'y a pas de différence sensible entre des personnes situées à des distances de 3, 4 voire appartenant à une autre composante connexe : il s'agit dans tous les cas de personnes avec lequelles ils ne partagent aucune connaissance

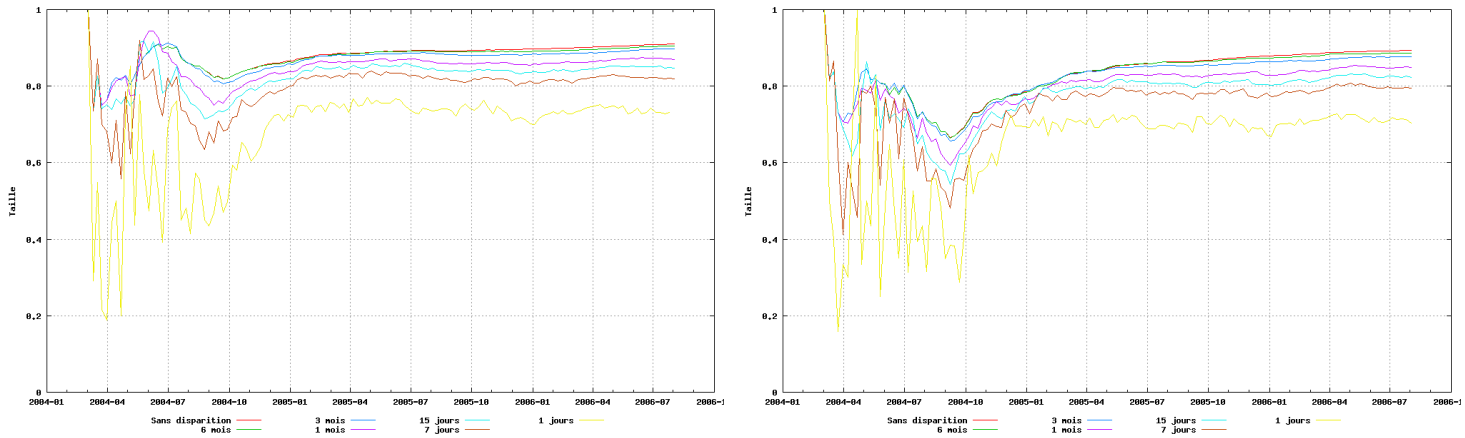


FIG. 4 – Taille relative de la composante connexe principale pour l'ensemble des liens (à gauche) et pour les liens réciproques (à droite).

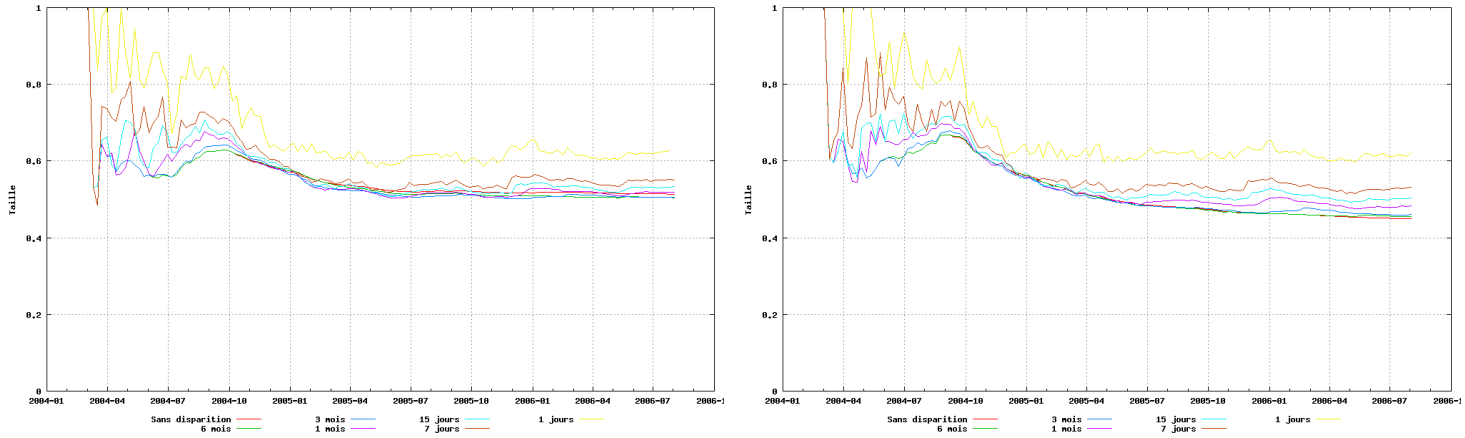


FIG. 5 – Taille relative de la périphérie pour l'ensemble des liens (à gauche) et pour les liens réciproques (à droite).

commune (voir [RC08]).

Liens répétés	Écarts de 2	Écarts ≥ 3	Total des commentaires
29 946 674	6 781 686	2 865 797	39 594 157
75,6%	17,2%	7,2%	100%

TAB. 1 – Répartition des commentaires en fonction du type de contact.

La table 1 indique les résultats de ces mesures. Les liens répétés représentent 75,6% de l'ensemble des commentaires et si on ajoute à cela les 17,2% de nouvelles relations avec des voisins proches, on constate que les commentaires sont échangés dans près de 93% des cas entre des utilisateurs qui sont déjà voisins ou qui ont au moins un voisin en commun. Une grande majorité de l'activité du réseau s'effectue donc sur de très courtes distances.

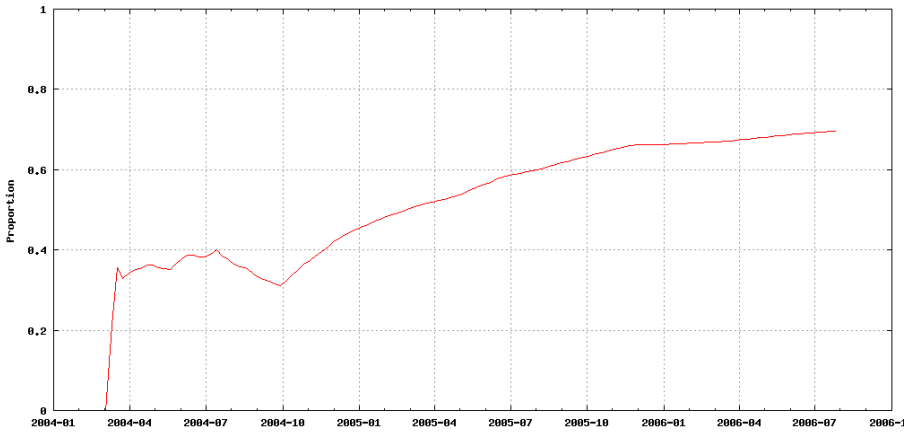


FIG. 6 – Évolution de la proportion d'écarts de 2 par rapport à l'ensemble des nouveaux liens.

On observe la même tendance si l'on s'intéresse seulement aux nouvelles relations : elles concernent dans 70% des cas des utilisateurs qui ont déjà au moins un contact en commun. La figure 6 représente l'évolution de cette proportion sur l'ensemble de la période. On distingue trois étapes successives : d'abord une forte augmentation avec un pic de 40%, puis une baisse qui conduit à un creux fin septembre 2004. La proportion ne cesse ensuite de croître pour le reste de la période. On retrouve ici les trois étapes décrites par Kumar, Novak et Tomkins à propos de l'évolution de la densité [KNT06]. Ce n'est pas surprenant dans la mesure où un nouveau contact entre deux sommets à distance 2 entraine la création d'un nouveau triangle, ce qui augmente la densité du réseau local. La proportion d'écarts de 2 parmi les nouveaux liens peut donc être considérée comme un bon indicateur de la densité du graphe.

Le réseau des commentaires se constitue donc en grande partie à travers des liens courts, le plus souvent au sein du voisinage ou du voisinage à distance 2 des utilisateurs. Cette tendance augmente tout au long de la période, ce qui entraine une densification du graphe.

Ces informations nous permettent de mieux comprendre la structure du réseau des commentaires, mais elles ne proposent que des tendances d'ordre général : les valeurs moyennes obtenues sont en effet écrasées par le volume des commentaires, et ne peuvent pas rendre compte de la diversité des situations que l'on peut rencontrer au sein du réseau. Il est donc nécessaire de s'intéresser aux parcours individuels pour s'affranchir des effets de masse et mieux comprendre la variété des comportements.

3 Parcours individuels et profils de sommets

3.1 La composition de l'entourage relationnel

Notre objectif est de caractériser différents types de nœuds et de comprendre comment ils construisent et organisent leur « réseau » de contacts. Une première étape consiste à faire des mesures pour chaque noeud du graphe afin de déterminer une liste d'indicateurs qui nous permettent de les comparer entre eux. Les informations retenues pour chaque sommet u sont le degré ($\deg(u)$) et la proportion de voisins proches ($P(u)$).

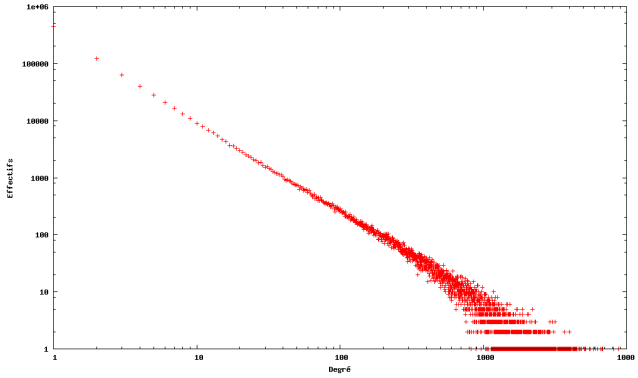


FIG. 7 – Distribution des degrés pour l'ensemble des commentaires.

Les figures 7 et 8 montrent la distribution de ces deux indicateurs. Comme pour tous les graphes de terrain (dont les grands réseaux sociaux), la distribution des degrés est très hétérogène, il y a un très grand nombre de sommets de très faible degré (438 840 sommets de degré 1) côtoyant quelques sommets avec un degré très élevé (le degré maximum est de 8 731).

Comme un sommet ne peut pas avoir de voisin proche s'il ne possède pas déjà un voisin pour servir d'intermédiaire, la distribution de la proportion de

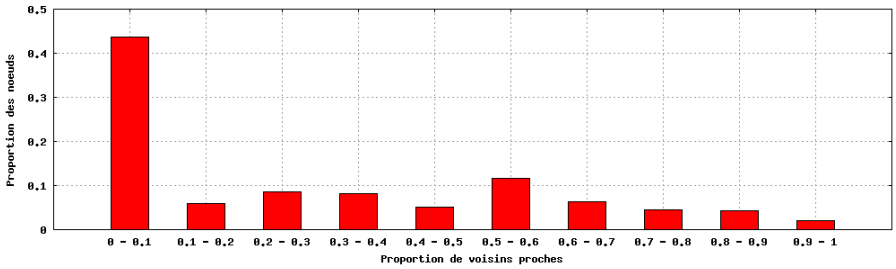
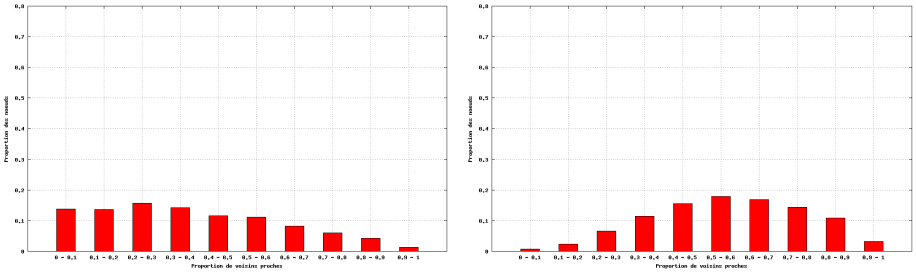
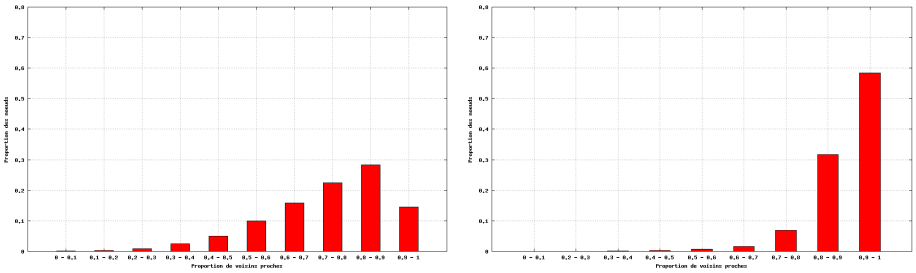


FIG. 8 – Distribution des proportions de voisins proches pour l'ensemble des commentaires.

voisins proches ne concerne que les sommets qui ont au minimum 2 voisins. Cette distribution est par ailleurs difficile à apprécier car elle est en partie biaisée par la distribution des valeurs possibles. Par exemple, les sommets de degré 2 ne peuvent avoir que deux valeurs pour $P(u)$, qui sont 0% s'ils n'ont aucun voisin proche et 50% s'ils en ont un. Comme les sommets de faible degré sont les plus nombreux, cela conduit à une forte proportion de sommets pour lesquels $P(u)$ prend une valeur de 0% ou de 50%, ce qui explique en grande partie les deux pics que l'on observe sur l'histogramme pour les classes [0%–10%] et [50%–60%]. Si l'on écarte ces deux pics, on constate que l'intervalle [10%–49%] est plus peuplé que l'intervalle [60%–100%].



Degrés 10 à 39. Degrés 40 à 99.



Degrés 100 à 749. Degrés $\geq$ à 750.

FIG. 9 – Distributions des proportions de voisins proches en fonction des degrés.

Le graphe des commentaires semble donc être composé d'une majorité de sommets qui ont une faible proportion de voisins proches, ce qui semble contradictoire avec la proportion moyenne de 70% de voisins proches que nous avons calculée pour l'ensemble du graphe. Pour résoudre ce problème, il faut comparer les distributions de $P(u)$ en regroupant les sommets en fonction de leur degré. La figure 9 représente les distributions obtenues. Plus le degré est important, plus les valeurs de $P(u)$ sont élevées : les sommets de plus faible degré ont des proportions relativements faibles, les sommets de degré intermédiaire ont une distribution homogène, et la proportion de voisins proches devient de plus en plus élevée pour les sommets de fort degré. Près de 90% des sommets dont le degré est supérieur ou égal à 750 ont une proportion de voisins proches supérieure à 80%.

L'interprétation de ces résultats est délicate, car la taille du voisinage à distance 2 augmente très fortement en fonction du degré : on peut supposer que les sommets de fort degré ont plus de chances d'entrer en contact avec un voisin proche, même s'il est choisi de façon aléatoire. On constate en effet que les sommets de degré 20 ont en moyenne 6 700 voisins à distance 2, avec un maximum de 25 700. Ces nombres sont élevés mais ils représentent malgré tout une très faible proportion de l'ensemble des sommets. Par ailleurs, tous les sommets de degré supérieur à 6 000 ont plus de 230 000 voisins à distance 2 soit environ un quart de l'ensemble des sommets, mais leur proportion de voisins proches dépasse les 90%. La très grande taille du voisinage à distance 2 pour les sommets de fort degré ne suffit donc pas à expliquer ce phénomène. On peut alors imaginer que les variations de la proportion de voisins proches pourraient refléter différentes pratiques dans la construction du réseau personnel.

On observe des résultats semblables si l'on effectue les mêmes mesures sur le graphe des commentaires réciproques, avec des ordres de grandeur réduits en raison de sa taille plus petite. Le degré maximum est alors de 2 930, et on peut obtenir des distributions de la proportion de voisins proches presques identiques en choisissant respectivement des classes de degré 10 à 29, 30 à 79, 80 à 499 et supérieur à 500.

3.2 Suivi individuel des sommets

Pour mieux comprendre les implications de ces pratiques, nous avons choisi de comparer l'évolution dans le graphe des liens réciproques de l'entourage d'un sommet qui privilégie les contacts lointains et d'un autre qui privilégie les contacts proches.

Les deux sommets sont de degré 80 : cette valeur est suffisament élevée pour que l'on puisse observer une réelle évolution, tout en restant dans une tranche de degrés où les proportions de contacts proches sont réparties de manière relativement homogène. Comme les contacts sont réciproques, nous savons que les degrés correspondent au même type de relations, ce qui ne serait pas le cas si nous choisissions de comparer deux sommets de même degré dans le graphe de l'ensemble des commentaires où deux sommets de même degré peuvent avoir

des types de relations très différents si un sommet a construit son entourage en envoyant de nombreux messages sans réponse, ou si au contraire il a reçu de nombreux messages auxquels il n'a pas nécéssairement répondu. Enfin, les deux sommets choisis sont tous deux actifs à partir de fin mars 2006.

Les figures 10 et 11 ont été obtenues en calculant pour chaque sommet l'évolution de son entourage, à intervalles d'une semaine (cf. partie 1.4). La première courbe montre l'évolution de son degré et de la proportion de ses voisins proches, qui apparaîssent en vert. Nous mesurons dans la deuxième courbe l'apparition de nouveaux voisins, avec là aussi la proportion de nouveaux voisins proches. La troisième courbe représente l'activité de notre individu : les messages émis apparaîssent en rouge et les messages reçus en vert. La dernière courbe représente l'évolution du nombre de ses voisins à distance 2.

Dans le cas du sommet A, qui a une faible proportion de voisins proches, le degré augmente par à-coups, avec des paliers successifs. Il a une activité équilibrée entre messages émis et reçus, dans l'ensemble inférieure à une dizaine de messages par semaine en dehors des pics ponctuels d'activité. Son voisinage à distance 2 grandit en même temps que son degré, on retrouve les mêmes paliers dans son évolution, dans un ordre de grandeur bien sûr largement supérieur puisque le sommet finit avec plus de 9 000 voisins à distance 2. Le sommet B favorise au contraire les contacts avec son voisinage proche : il est plus actif, avec une vingtaine de messages émis et reçus par semaine en moyenne et un pic d'activité de plus 50 messages. Son voisinage augmente de façon régulière tout au long de la période, mais il ne dépasse pas les 2 000 voisins à distance 2.

3.3 La structure des voisinages

L'évolution différente du degré des deux sommets peut s'expliquer par la différence d'effort qu'il faut mobiliser pour établir un contact. En effet, un utilisateur peut voir immédiatement les commentaires déposés par les voisins de ses voisins en consultant les photos de ses voisins, il n'a aucune recherche à effectuer avant d'entrer en contact avec eux.

Le sommet B a une activité plus régulière et plus intense que le sommet A. On peut imaginer que cette activité plus intense lui permet d'avoir une meilleure connaissance des voisins de ses voisins et qu'il peut plus facilement établir de nouvelles relations avec eux. Le sommet A s'investit moins dans le service, sauf lors de pics d'activité qui peuvent être interprétés comme des périodes d'intérêt ponctuel pendant lesquelles il entre en contact avec de nouveaux voisins. Comme il entretient moins de contacts avec ses anciens voisins, on peut supposer qu'il connaît moins bien leur voisinage et a donc moins de chances d'y découvrir des sommets susceptibles de l'intéresser.

Ces différents modes de construction de l'entourage pourraient expliquer l'écart important de la taille des voisinages à distance 2. Lorsqu'un sommet entre en contact avec un voisin proche, il y a une forte probabilité pour qu'une partie du voisinage du nouveau contact appartienne déjà au voisinage du sommet. Ce type de relations va densifier le voisinage direct du sommet, mais apporte peu de nouveaux sommets dans son voisinage à distance 2. Lorsqu'un sommet entre en

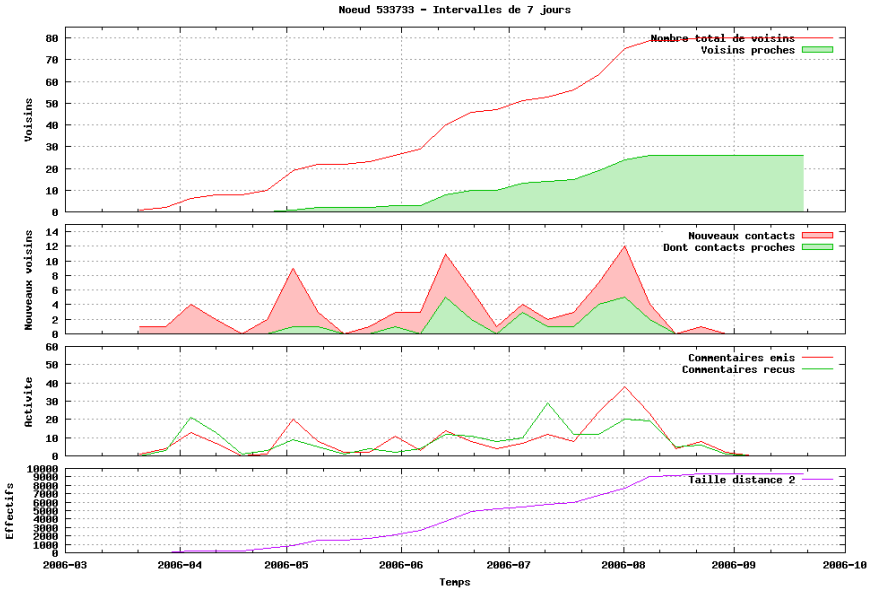


FIG. 10 – Évolution de l'entourage du sommet A, qui privilégie les contacts lointains.

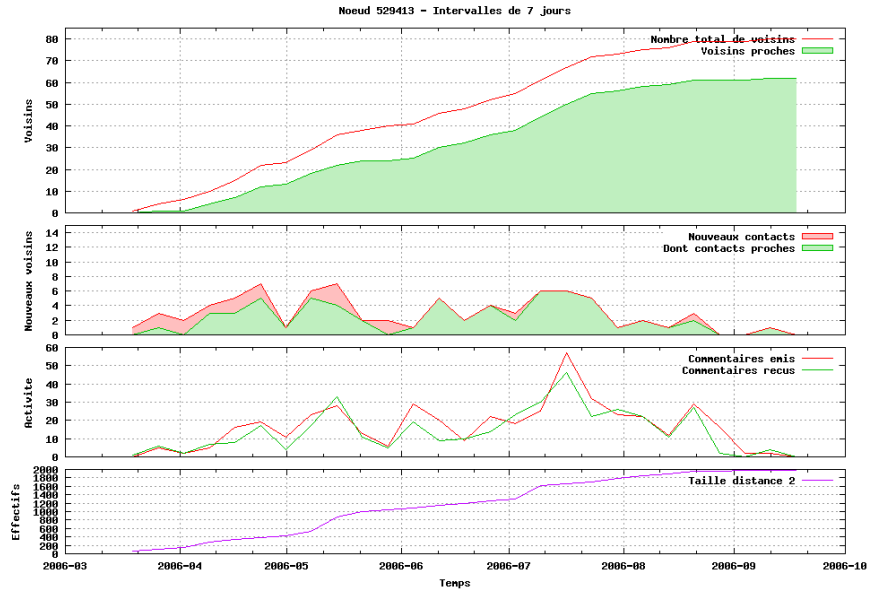


FIG. 11 – Évolution de l'entourage du sommet B, qui privilégie les contacts proches.

Les échelles sont identiques pour les deux figures, sauf pour l'évolution de la taille de la distance 2 : la figure 10 a un maximum de 10 000 tandis que la figure 11 a un maximum de 2 000 seulement.

contact avec des sommets plus lointains, leurs voisinages respectifs ont moins de chances de se chevaucher et la taille du voisinage à distance 2 a plus de chance d'augmenter sensiblement.

On peut imaginer que ces deux comportements ont des conséquences sur la structure de l'entourage : le fait de favoriser les contacts proches entrainerait une densification de l'entourage, mais au prix d'une plus faible ouverture sur le reste du graphe, ce qui favoriserait l'émergence de cliques communautaires, *i.e.* de réseaux locaux très denses et peu liés avec le reste du graphe. Les contacts répétés avec des sommets lointains favoriseraient au contraire une diversification du réseau local dans lequel le sommet tiendrait un rôle plus central, puisque c'est par son intermédiaire que les nouvelles relations au sein de son entourage vont s'établir[1].

Bien sûr, ces deux exemples ne sont pas représentatifs de la variété des situations que l'on peut observer, mais l'analyse de ces cas particuliers nous permet de mieux comprendre les mécanismes de construction des entourages et nous pousse à imaginer de nouvelles méthodes pour répondre aux questions qu'ils suscitent. Nous pouvons ainsi mesurer le rythme de l'activité des sommets dans le temps pour mieux comprendre les différents comportements.

4 Gestion du temps par l'individu

4.1 Mesure de l'activité de chaque individu

Pour mesurer les rythmes de l'activité acteurs du réseau, nous introduisons quelques mesures supplémentaires. Nous déclarons un *sommet actif* au cours d'un intervalle de temps donné s'il a émis au moins un commentaire au cours de cet intervalle. La *durée d'activité d'un lien* est l'intervalle de temps écoulé entre sa première et sa dernière occurrence. De la même manière, la *durée d'activité d'un sommet* est l'intervalle de temps écoulé entre les dates de première et dernière activité du sommet. Ces mesures ne tiennent pas compte des variations dans l'activité des sommets et en particulier des éventuelles périodes d'inactivité : nous choisissons pour cela de découper T en intervalles de temps fixe, et nous définissons *l'activité réelle d'un sommet* comptant le nombre d'intervalles au cours desquels le sommet est actif. Nous avons choisi pour notre étude d'exprimer l'activité réelle en intervalles d'une semaine.

La figure 12 montre la distribution des durées d'activité, exprimées en jours, pour l'ensemble des sommets du graphe. On dénombre 567 477 sommets avec une durée d'activité inférieure à une journée, et la durée maximale est de 925 jours. La distribution semble décroître faiblement[2] pour les sommets qui ont une durée comprise entre 100 et 500 jours. Si l'on s'intéresse à l'activité des sommets

[1] Dans le vocabulaire de l'analyse des réseaux sociaux, on parlerait de *trous structuraux* [Bur92].

[2] L'utilisation d'une échelle logarithmique pour l'axe des coordonnées ne permet pas de voir en détail cette partie.

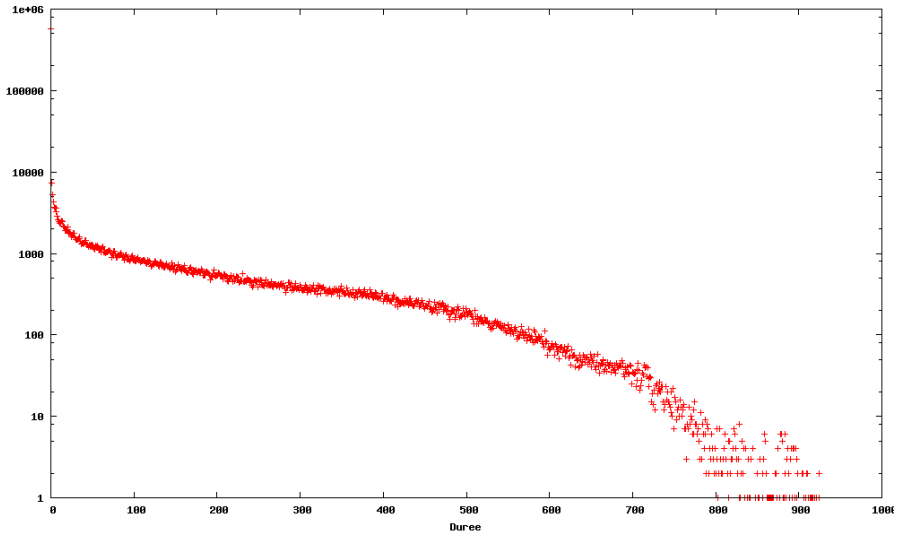


FIG. 12 – Distribution des durées d'activité des sommets pour l'ensemble des commentaires.

pour le graphe des commentaires réciproques, on obtient une distribution similaire, mais avec une proportion beaucoup moins importante de sommets dont la durée est inférieure à une journée, que nous désignerons comme des *sommets éphémères*. Nous ne présentons pas ici la distribution de l'activité des liens, mais elle suit une distribution tout à fait comparable, avec un grand nombre de liens de très courte durée, un intervalle de durées relativement stable, et quelques liens de longue durée. La distribution de l'activité réelle des sommets est quant à elle plus hétérogène que celle des durées d'activité, tout en restant dans les mêmes ordres de grandeur.

Durée d'activité	Tous les commentaires	Commentaires réciproques
moins d'une journée	$62,3\%$	$26,7\%$
moins d'une semaine	$65,3\%$	$30,8\%$
moins d'un mois	$70,7\%$	$39,3\%$
moins de 3 mois	$78,3\%$	$53,2\%$
moins de 6 mois	$85,5\%$	$67,5\%$
moins d'un an	$94,1\%$	$86,5\%$

TAB. 2 – Répartition des proportions de sommets actifs en fonction de la durée d'activité.

Nous comparons les resultats des mesures sur les deux graphes dans le tableau 2 : les sommets actifs pendant moins d'une journée représentent $62,3\%$ de l'ensemble des sommets si on considère l'ensemble des commentaires, mais

seulement 26, 7% pour le graphe des commentaires réciproques. Cette différence n'est pas surprenante puisque nous savons que par construction, tout commentaire a forcément reçu une réponse dans le graphe des commentaires réciproques. Cependant en pratique, une grande partie de ces réponses arrive très rapidement après l'envoi du premier message (voir [KNT06]). L'écart important des proportions, qui perdure pour des durées d'une semaine ou d'un mois nous confirme que le graphe des commentaires réciproques s'affranchit d'un grand nombre d'utilisateurs ponctuels, même si leur proportion reste importante.

4.2 Concentration de l'activité

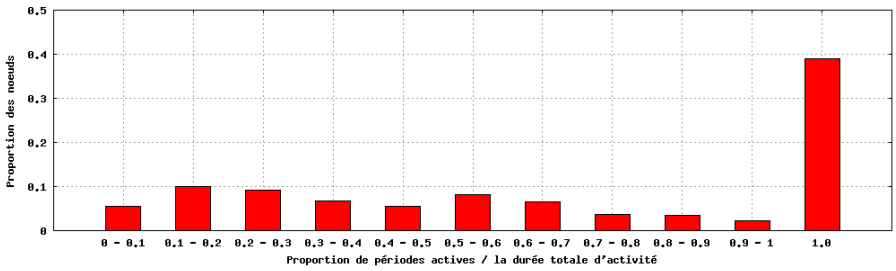
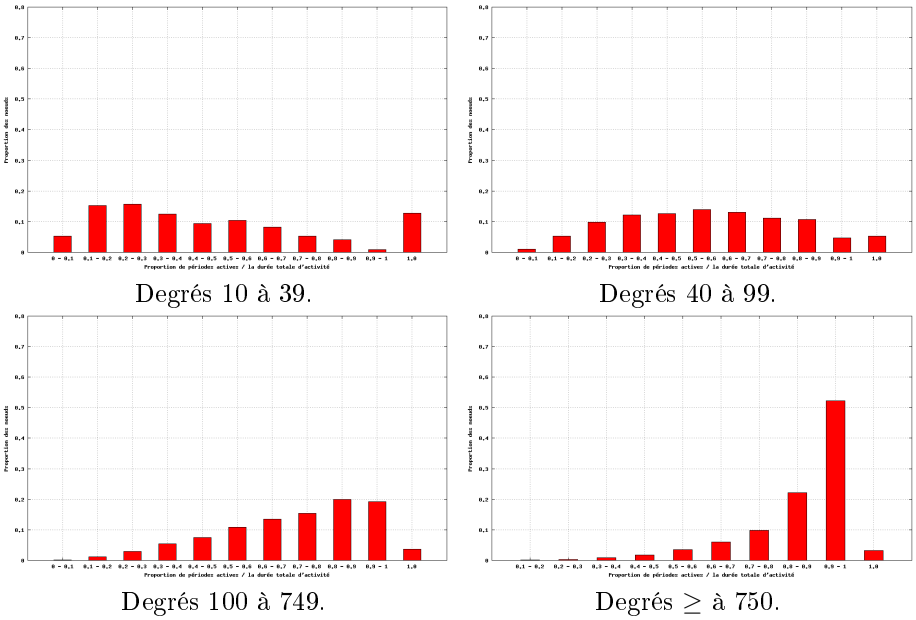


FIG. 13 – Distribution de la concentration de l'activité des sommets par tranches d'une semaine.

On peut alors mesurer la *concentration de l'activité* d'un sommet en faisant le rapport entre son activité réelle et sa durée d'activité : nous calculons donc pour chaque sommet le rapport entre le nombre de semaines pendant lesquelles il a émis au moins un commentaire et le nombre de semaines écoulées entre son premier et son dernier commentaire. La figure 13 représente la distribution de la concentration de l'activité des sommets pour l'ensemble des commentaires. On remarque la présence d'un pic de sommets qui ont une concentration de 100% : cela correspond aux utilisateurs éphémères, qui se désintéressent rapidement du service et l'abandonnent dès la première semaine.

La figure 14 représente les variations de cette distribution selon quatre classes de degré. On constate que les distributions ressemblent à celles de la proportion de voisins proches. Les sommets de faible degré ont tendance à avoir une activité plus diffuse, malgré la présence d'un pic de concentration à 100% qui correspond probablement à des utilisateurs ponctuels. Les sommets de degré intermédiaire ont une distribution homogène, et la concentration augmente ensuite de plus en plus pour les sommets de fort degré.

Comme notre mesure ne concerne que les messages émis, les concentrations élevées pour les sommets de fort degré ne sont pas une conséquence de la taille de leur entourage : plus un sommet a de voisins et plus il a de chances de recevoir un message de l'un d'entre eux lors de chaque période. Lorsque la concentration est faible, cela signifie que le sommet a une activité diluée dans le temps, ce qui peut correspondre à des utilisateurs qui utilisent le service de manière occasionnelle.



Degrés 10 à 39. Degrés 40 à 99.

Degrés 100 à 749. Degrés $\geq$ à 750.

FIG. 14 – Distributions de la concentration de l'activité en fonction du degré.

Les distributions obtenues pour les sommets de fort degré montrent qu'une forte proportion d'entre eux se caractérise par une activité continue, avec très peu de périodes d'interruption.

Références

[BA99] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 1999.

[Boy04] Danah Boyd. Friendster and publicly articulated social networks. In *Conference on Human Factors and Computing Systems (CHI 2004)*, 2004.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7) :107–117, 1998.

[BS03] S. Bornholdt and H. G. Schuster, editors. *Handbook of Graphs and Networks*. Wiley-Vch, 2003.

[Bur92] R. Burt. *Structural Holes. The Social Structure of Competition*. Cambridge, Harvard University Press, 1992.

[CP08] Dominique Cardon and Christophe Prieur. *Les réseaux de relations sur Internet : un objet de recherche pour l'informatique et les sciences sociales*, chapter in Humanités numériques, C. Brossaud and B. Reber (ed.). Hermès, 2008.

[DF94] A. Degenne and M. Forse. *Les réseaux sociaux*. Armand Colin, 1994.

[Eli91] Norbert Elias. *La société des individus*. Fayard, 1991.

[ERC+00] Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, and Seyda Ertekin. The shape of the Web and its implications for searching the Web, 2000.

[Gra78] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, pages 1360–1380, 1978.

[Gri98] Maurizio Gribaudi, editor. *Espaces, temporalités, stratifications : Exercices sur les réseaux sociaux*. éditions de l'EHESS, 1998.

[JGN01] Emily M. Jin, Michelle Girvan, and M. E. J. Newman. The structure of growing social networks. Working Papers 01-06-032, Santa Fe Institute, June 2001.

[KNT06] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[Lat08] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science (TCS)*, 407 :458–473, 2008.

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 167(45), 2003.

[O'R05] Tim O'Reilly. What is web 2.0 : design patterns and business models for the next generation of software. http ://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, 2005.

[PBV07] Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution. *Nature*, 446(7136) :664–667, 2007.

[PCB+09] Christophe Prieur, Dominique Cardon, Jean-Samuel Beuscart, Nicolas Pissard, and Pascal Pons. La photo comme conversation : une étude de cas sur flickr. *Réseaux*, 2009.

[PSS09] Christophe Prieur, Alina Stoica, and Zbigniew Smoreda. Extraction de réseaux égocentrés dans un (très grand) réseau social. *Bull. de Méthodologie Sociol.*, 101, 2009.

[RC08] Camille Roth and Jean-Philippe Cointet. Social and semantic coevolution in knowledge networks. *Social Networks*, special issue on the Dynamics of Social Networks, 2008.

[Sco92] John Scott. *Social Network Analysis*. Sage, London, 1992.

[Sei83] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 1983.

[WD07] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34 :441–458, 2007.

[Wel93] Barry Wellman. An egocentric network tale. *Social Networks*, 15 :423–436, 1993.

[WS98] Duncan Watts and Steve Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 1998.